Online Forecasting of Total-Variation-bounded Sequences

Dheeraj Baby [1] **Yu-Xiang Wang** [1]

Abstract

We consider the problem of online forecasting of sequences of length n with total-variation at most C_n using observations contaminated by independent σ-subgaussian noise. We design an $O(n \log n)$-time algorithm that achieves a cumulative square error of $\tilde{O}(n^{1/3} C_n^{2/3} \sigma^{4/3})$ with high probability. The result is rate-optimal as it matches the known minimax rate for the offline nonparametric estimation of the same class (Mammen & van de Geer, 1997). We show that online gradient descent and its variants with a fixed restarting schedule — are instances of a class of *linear forecasters* that require a suboptimal regret of $\tilde{\Omega}(\sqrt{n})$. This implies that the use of more complex algorithms are necessary to obtain the optimal rate. To the best of our knowledge, this is the first work of its kind that considers local adaptivity in an online forecasting problem, and the first non-trivial class of online learning problems with a minimax *dynamic* regret of $O(n^{1/3})$.

1. Introduction

Nonparametric regression is a fundamental class of problems that has been studied for more than half a century in statistics and machine learning (Nadaraya, 1964; De Boor et al., 1978; Wahba, 1990; Donoho et al., 1998; Mallat, 1999; Scholkopf & Smola, 2001; Rasmussen & Williams, 2006). It solves the following problem:

- Let $y_i = f(x_i) + \text{Noise}$ for $i = 1, ..., n$. How can we estimate a function f using data points $(x_1, y_1), ..., (x_n, y_n)$ in conjunction with the knowledge that f belongs to a function class $\mathcal{F}$?

A recent and successful class of nonparametric regression technique called trend filtering (Steidl et al., 2006; Kim et al., 2009; Tibshirani, 2014; Wang et al., 2014) was shown to have the property of *local adaptivity* (Mammen & van de Geer, 1997) in both theory and practice. We say a nonparametric regression technique is *locally adaptive* if it can cater to local differences in smoothness, hence allowing more accurate estimation of functions with varying smoothness and abrupt changes. For example, for functions with bounded total variation (when $\mathcal{F}$ is a total variation class), standard nonparametric regression techniques such as kernel smoothing and smoothing splines have a mean square error (MSE) of $O(n^{-1/2})$ while trend filtering has the optimal $O(n^{-2/3})$.

Trend filtering is, however, a batch learning algorithm where one observes the entire dataset ahead of the time and makes inference about the past. This makes it inapplicable to the many time series problems that motivate the study of trend filtering in the first place (Kim et al., 2009). The focus of this work is in developing theory and algorithms for locally adaptive online forecasting which predicts the immediate future value of a function with heterogeneous smoothness using only noisy observations from the past.

1.1. Problem Setup

We propose a model for nonparametric online forecasting as described in Figure 1. This model can be re-framed in the language of the online convex optimization model with three differences.

1. We consider only quadratic loss functions of the form $f_t(x) = (x - \theta_t)^2$.
2. The learner receives independent *noisy* gradient feedback, rather than the exact gradient.
3. The criterion of interest is redefined as the *dynamic regret* (Zinkevich, 2003; Besbes et al., 2015):

$$R_{\text{dynamic}}(\hat{\theta}, f_{1:n}) := \mathbb{E}\left[\sum_{t=1}^{n} f_t(x_t)\right] - \sum_{t=1}^{n} \inf_{x_t \in \Theta} f_t(x_t). \tag{1}$$

1.2. Assumptions

We consolidate all the assumptions used in this work and provide necessary justifications for them.

[*] Equal contribution [1] Department of Computer Science, University of California, Santa Barbara. Correspondence to: Dheeraj Baby <dheeraj@ucsb.edu>, Yu-Xiang Wang <yuxiangw@cs.ucsb.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

1. Fix action time intervals $1, 2, ..., n$
2. The player declare a forecasting strategy $\hat{f}_i : \mathbb{R}^{i-1} \to \mathbb{R}$ for $i = 1, ..., n$.
3. An adversary θ_i for $i = 1, ..., n$.
4. For every time point $i = 1, ..., n$:
   (a) We play $a_i = \hat{f}_i(y_1, ..., y_{i-1})$.
   (b) We receive a feedback $y_i = \theta_i + Z_i$, where Z_i is a zero-mean, independent subgaussian noise.
5. At the end, the player suffers a cumulative regret of $\sum_{i=1}^{n} (a_i - \theta_i)^2$.

Figure 1. Nonparametric online forecasting model. The focus of the proposed work is to design a forecasting strategy that minimizes the expected cumulative regret. Note that the problem depends a lot on the choice of the sequence θ_i. Our interest is on sequences with bounded total variation (TV) so that $\sum_{i=2}^{n} |\theta_i - \theta_{i-1}| \leq C_n$.

- (A1) The time horizon for the online learner is known to be n.
- (A2) The parameter σ^2 of subgaussian noise in the observations is known.
- (A3) The sequence $\boldsymbol{\theta} = [\theta_1, ..., \theta_n]$ has its total variation bounded by some known positive C_n, i.e., we take $\mathcal{F}$ to be the total variation class $\mathrm{TV}(C_n) := \{\boldsymbol{\theta} \in \mathbb{R}^n | \|D\boldsymbol{\theta}\|_1 \leq C_n\}$ where D is the discrete difference operator.
- (A4) $|\theta_1| \leq U$.

The knowledge of σ^2 in assumption (A2) is primarily used to get the optimal dependence of σ in minimax rate. This assumption can be relaxed in practice by using the Median Absolute Deviation estimator as described in Section 7.5 of Johnstone (2017) to estimate σ^2 robustly. Assumption (A3) features a large class of functions with spatially inhomogeneous degree of smoothness. The functions residing in this class need not even be continuous. Our goal is to propose a policy that is locally adaptive whose empirical mean squared error converges at the minimax rate for this function class. The knowledge of C_n is used to get its optimal dependence on the regret. Assumption (A4) is very mild as it puts restriction only to the first value of the sequence. This assumption controls the inevitable prediction error for the first point in the sequence.

1.3. Our Results

Contributions The major contributions of this work are summarized below.

- It is known that the minimax MSE for *smoothing* sequences in the TV class is $\tilde{\Omega}(n^{-2/3})$. This implies a lowerbound of $\tilde{\Omega}(n^{1/3})$ for the dynamic regret in our setting. We present a policy ARROWS (**A**daptive **R**estarting **R**ule for **O**nline averaging using **W**avelet **S**hrinkage) with a nearly minimax dynamic regret $\tilde{O}(n^{1/3})$.
- We show that a class of forecasting strategies — including the popular Online Gradient Descent (OGD) with fixed restarts and moving averages — are fundamentally limited by $\tilde{\Omega}(\sqrt{n})$ regret.
- We also provide a more refined lower bound that characterized the dependence of U, C_n and σ, which certifies the optimality of ARROWS in all regimes. The bound also reveals a subtle price to pay when we move from the smoothing problem to the forecasting problem, which indicates the separation of the two problems when $C_n/\sigma \gg n^{1/4}$, a regime where the forecasting problem is *strictly* harder (See Figure 3).
- Naive implementation of our policy will have a runtime complexity of $O(n^2)$. We exploit the sequential structure of our policy and sparsity in wavelet transforms to construct an $O(n \log(n))$ implementation.

1.4. A brief comparison to existing online non-parametric methods

We note that our problem falls into the more general framework of online non-parametric regression setting studied in (Rakhlin & Sridharan, 2015). It can be shown that our dynamic regret minimization setting is reducible to theirs. Since the bounded TV class is sandwiched between Besov spaces $B^1_{1,q}$ for the range $1 \leq q \leq \infty$, the discussion in section 5.8 of (Rakhlin & Sridharan, 2015) establishes that minimax growth w.r.t n as $O(n^{1/3})$ in the online setting for TV class. Thus our bound, modulo logarithmic factor, matches with theirs though we give the precise dependence on C_n and σ as well. It is worthwhile to point out that while the bound in (Rakhlin & Sridharan, 2015) is non-constructive, we achieve the same bound via an efficient algorithm.

2. Main results

We are now ready to present our main results.

2.1. Limitations of Linear Forecasters

Let's consider the class of linear forecasters — estimators that outputs a fixed linear transformation of the observations $y_{1:n}$. The following preliminary result shows that Restarting OGD (Besbes et al., 2015; Chen et al., 2018) is a linear forecaster . By the results of Donoho et al. (1998), linear smoothers are fundamentally limited in their ability to estimate functions with heterogeneous smoothness. Since forecasting is harder than smoothing, this limitation gets directly translated to the setting of linear forecasters.

Proposition 1. *Online gradient descent with a fixed restart schedule is a linear forecaster. Therefore, it has a dynamic regret of at least $\tilde{\Omega}(\sqrt{n})$.*

The proposition implies that we will need fundamentally new *nonlinear* algorithmic components to achieve the optimal $O(n^{1/3})$ regret, if it is achievable at all!

2.2. Policy

In this section, we present our policy ARROWS (Adaptive Restarting Rule for Online averaging using Wavelet Shrinkage). The following notations are introduced for describing the algorithm.

- t_h denotes start time of the current bin and t be the current time point
- $\bar{y}_{t_h:t}$ denotes the average of the y values for time steps indexed from t_h to t.
- $pad_0(y_{t_h}, ..., y_t)$ denotes the vector $(y_{t_h} - \bar{y}_{t_h:t}, ..., y_t - \bar{y}_{t_h:t})^T$ zero-padded at the end till its length is a power of 2. *i.e*, a re-centered and padded version of observations.
- $T(x)$ where x is a sequence of values, denotes the element-wise soft thresholding of the sequence with threshold $\sigma\sqrt{\beta \log(k)}$ where k is the length of x.
- H denotes the orthogonal discrete Haar wavelet transform matrix of proper dimensions

2.3. Dynamic Regret of ARROWS

In this section, we bound the run-time and non-stationary regret of the policy.

Theorem 1. *Let the feedback be $y_t = \theta_t + Z_t$ where Z_t is an independent, σ-subgaussian random variable. Let $\theta_{1:n} \in \mathrm{TV}(C_n)$. If $\beta = 6 + \frac{2\log(8/\delta)}{\log(n)}$, then with probability at least $1 - \delta$,* ARROWS *achieves a dynamic regret of $\tilde{O}(n^{1/3}C_n^{2/3}\sigma^{4/3} + U^2 + C_n^2 + \sigma^2)$ where $\tilde{O}$ hides a logarithmic factor in n and $1/\delta$*

Remark 1. By a rewriting of the regret bound, it can be shown that our policy is also optimal for predicting sequences from Sobolev space defined by sequences that satisfy $\|D\theta(1:n)\|_2 \leq C'_n = n^{-1/2}C_n$. In other words, our policy is adaptively minimax, adaptive to the underlying function class being Sobolev or TV class.

```
ARROWS: inputs - observed $y$ values, time horizon $n$, $\delta \in (0,1]$, total variation bound $C_n$, a hyper-parameter $\beta > 6$
  1. Initialize $t_h = 1$, $newBin = 1$, $y_0 = 0$
  2. For $t$ = 1 to $n$:
     (a) if $newBin == 1$, predict $x_t^{t_h} = y_{t-1}$, else predict $x_t^{t_h} = \bar{y}_{t_h:t-1}$
     (b) set $newBin = 0$, observe $y_t$ and suffer loss $(x_t^{t_h} - \theta_t)^2$
     (c) Let $\hat{y} = pad_0(y_{t_h}, ..., y_t)$ and $k$ be the padded length.
     (d) Let $\hat{\alpha}(t_h : t) = T(H\hat{y})$
     (e) Restart Rule: If $\frac{1}{\sqrt{k}}\sum_{l=0}^{\log_2(k)-1} 2^{l/2}\|\hat{\alpha}(t_h : t)[l]\|_1 > n^{-1/3}C_n^{1/3}\sigma^{2/3}$ then
         i. set $newBin = 1$
         ii. set $t_h = t + 1$
```

2.4. A lower bound on the minimax regret

We now prove a matching lower bound of the expected regret.

Proposition 2. *Assume $\min\{U, C_n\} > 2\pi\sigma$ and $n > 3$, there is a universal constant c such that*

$$\inf_{x_{1:n}} \sup_{\theta_{1:n} \in \mathrm{TV}(C_n)} \mathbb{E}\left[\sum_{t=1}^{n} (x_t(y_{1:t-1}) - \theta_t)^2\right] \geq c(U^2 + C_n^2 + \sigma^2 \log n + n^{1/3}C_n^{2/3}\sigma^{4/3}). \tag{2}$$

Remark 2 (The price of forecasting)**.** The lower bound implies that a term with C_n^2 is required even if $\sigma = 0$, whereas in the case of a one-step look-ahead oracle (or the smoothing algorithm that sees all n observations) does not have this term. This implies that the total amount of variation that the algorithm can handle while producing a sublinear regret has dropped from $C_n = o(n)$ to $C_n = o(\sqrt{n})$. Moreover, the regime where the $n^{1/3}C_n^{2/3}\sigma^{4/3}$ term is meaningful only when $C_n = o(n^{1/4})$. For the region where $n^{1/4} \ll C_n \ll n^{1/2}$, the minimax regret is essentially proportional to C_n^2. This is illustrated in Figure 3.

Remark 3. It is worth pointing out that knowledge of σ and C_n in the policy is primarily used to get the optimal dependence of $\sigma^{4/3}C_n^{2/3}$ in the minimax regret. One can still get a regret that grows as $n^{1/3}$ even without the knowledge of these parameters if it is in the achievable regime by taking them to be a fixed constant in the restart criterion.

2.5. Fast computation

Last but not least, we present a fast and practical implementation of our proposed algorithm, which reduces the overall

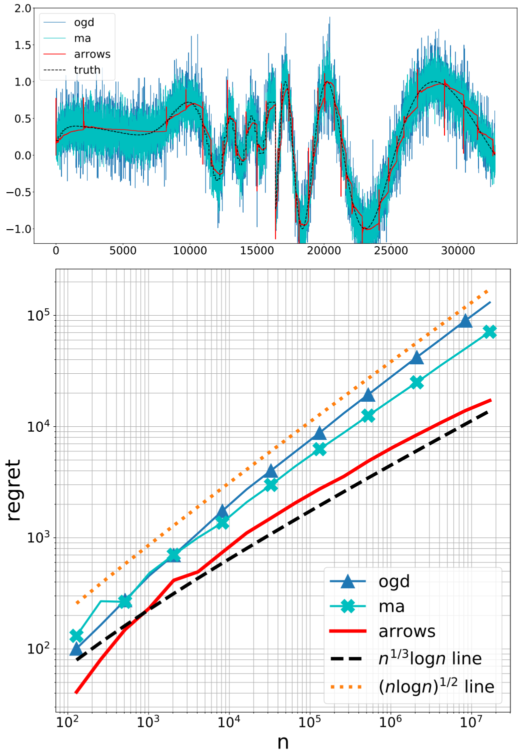

Figure 2. An illustration of ARROWS on a sequence with *heterogeneous smoothness*. We compare qualitatively (on the left) and quantitatively (on the right) to two popular baselines: (a) restarting online gradient descent (Besbes et al., 2015); (b) the moving averages (Box & Jenkins, 1970) with optimal parameter choices. As we can see, ARROWS achieves the optimal $n^{1/3}$ regret while the baselines are both suboptimal.

time-complexity for n step from a naive $O(n^2)$ algorithm to a nearly linear $O(n \log n)$ algorithm.

Proposition 3. *The run time of* ARROWS *is* $O(n\log(n))$, *where* n *is the time horizon.*

2.6. Demonstrating the adaptivity of ARROWS

Figure 2 shows the results on a function with heterogeneous smoothness with the hyperparameters selected according to their theoretical optimal choice for the TV class.

The top panel illustrates that ARROWS is locally adaptive to heterogeneous smoothness of the ground truth. Red peaks in the figure signifies restarts. During the initial and final duration, the signal varies smoothly and ARROWS chooses a larger window size for online averaging. In the middle, signal varies rather abruptly. Consequently ARROWS chooses a smaller window size. On the other hand, the linear smoothers OGD and MA use a constant width and cannot adapt to the different regions of the space. This differences are also reflected in the quantitative evaluation on the right, which clearly shows that OGD and MA has a suboptimal $\tilde{O}(\sqrt{n})$ regret while ARROWS attains the $\tilde{O}(n^{1/3})$ minimax regret!

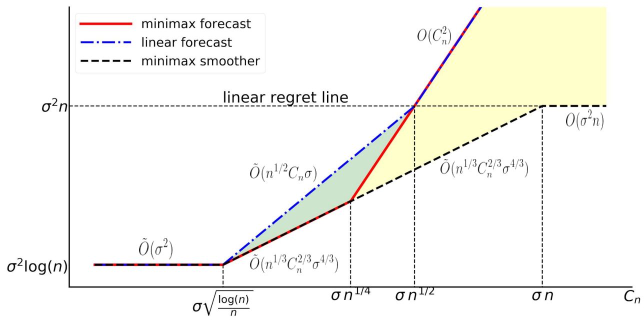

Figure 3. An illustration of the minimax (dynamic) regret of forecasters and smoothers as a function of C_n. The non-trivial regime for forecasting is when C_n lies between $\sigma\sqrt{\frac{\log(n)}{n}}$ and $\sigma\ n^{1/4}$ where forecasting is just as hard as smoothing. When $C_n > \sigma n^{1/4}$, forecasting is harder than smoothing. The yellow region indicates the extra loss incurred by any minimax forecaster. The green region marks the extra loss incurred by a linear forecaster compared to minimax forecasting strategy. The figure demonstrates that linear forecasters are sub-optimal even in the non-trivial regime. When $C_n > \sigma\ n^{1/2}$, it is impossible to design a forecasting strategy with sub-linear regret. For $C_n > \sigma\ n$, identity function is optimal estimator for smoothing and when when $C_n < \sigma^2\log(n)$, online averaging is optimal for both problems.

3. Concluding Discussion

In this paper, we studied the problem of forecasting bounded variation sequences. We proposed a new forecasting policy ARROWS, which we show to enjoy a dynamic regret of $\tilde{O}(n^{1/3}C_n^{2/3}\sigma^{4/3} + \sigma^2 + U^2 + C_n^2)$. We also derived a lowerbound which matches the upper bound up to a logarithmic term which certifies the optimality of ARROWS in all parameters. Further we exploited the sequential structure of our policy to devise an implementation with nearly linear run-time. Through the connection to linear estimation theory, we assert that that many existing online learners, such as those of Besbes et al. (2015) are linear forecasters and cannot achieve the optimal rate.

References

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Box, G. E. and Jenkins, G. M. *Time series analysis: forecasting and control*. John Wiley & Sons, 1970.

Chen, X., Wang, Y., and Wang, Y.-X. Non-stationary stochastic optimization under lp, q-variation measures. *Operations Research, to appear.*, 2018.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., and De Boor, C. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

Donoho, D. L., Johnstone, I. M., et al. Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3): 879–921, 1998.

Johnstone, I. M. *Gaussian estimation: Sequence and wavelet models*. 2017.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009.

Mallat, S. *A wavelet tour of signal processing*. Elsevier, 1999.

Mammen, E. and van de Geer, S. Locally apadtive regression splines. *Annals of Statistics*, 25(1):387–413, 1997.

Nadaraya, E. A. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

Rakhlin, A. and Sridharan, K. Online nonparametric regression with general loss functions, 2015.

Rasmussen, C. E. and Williams, C. K. *Gaussian processes for machine learning*. MIT Press, 2006.

Scholkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Steidl, G., Didas, S., and Neumann, J. Splines in higher order TV regularization. *International Journal of Computer Vision*, 70(3):214–255, 2006.

Tibshirani, R. J. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014.

Wahba, G. *Spline models for observational data*, volume 59. Siam, 1990.

Wang, Y.-X., Smola, A., and Tibshirani, R. The falling factorial basis and its statistical applications. In *International Conference on Machine Learning (ICML-14)*, pp. 730–738, 2014.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning (ICML-03)*, pp. 928–936, 2003.